STATISTICS WORKSHEET-1

1-a) True
2-c) Centroid Limit Theorem
3-b) Modeling bounded count data
4-c) The square of a standard normal random variable follows what is called chi-squared distribution
5-c) Poisson
6-b) False
7-b) Hypothesis
8-a) 0
9-c) Outliers cannot conform to the regression relationship.

10-The normal distribution is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "bell curve".

11-Imputation is the process of replacing missing values with substituted data. It is done as a preprocessing step.

Impute missing data:

- Mean/Median Imputation: In a mean or median substitution, the mean or a median value of a variable is used in place of the missing data value for that same variable.
- Mode substitution: In mode substitution, the highest occurring value for categorical value is used in place of the missing data value of the same variable.
- Substitution: Impute the value from a new individual who was not selected to be in the sample.
- Hot deck imputation
- Cold deck imputation
- Regression imputation
- Stochastic regression imputation
- Interpolation and extrapolation

12-A/B testing (also known as bucket testing, split-run testing, or split testing) is a user experience research methodology. A/B tests consist of a randomized experiment that usually involves two variants (A and B), although the concept can be also extended to multiple variants of the same variable. A/B testing is a way to compare multiple versions of a single variable, for example by testing a subject’s response to variant A against variant B, and determining which of the variants is more effective.

A/B testing is used by data engineers, marketers, designers, software engineers, and entrepreneurs, among others.

13-Mean imputation is one of the most used methods for handling missing data. However, it has some limitations. Mean imputation can lead to biased estimates if the data is not missing at random. It can also lead to an underestimation of the standard error of the estimate. Therefore, it is important to consider other methods of imputation such as multiple imputation.

14-Linear regression in statistics-Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:
(1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
(2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?
These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the

formula $y = c + b*x$,

where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.
Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressed. The independent variables can be called exogenous variables, predictor variables, or regressors.
Three major uses for regression analysis are
(1) determining the strength of predictors,
(2) forecasting an effect,
(3) trend forecasting.

15-Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data1. The two main branches of statistics are **descriptive statistics** and **inferential statistics**. Descriptive statistics deals with the presentation and collection of data. Inferential statistics is used to make inferences about a population based on a sample of data.